454 Sequencing Is an Effective Method for Gap Closure in Microbial Whole Genome Shotgun Sequencing

Feng Chen, Jamie Jett, Edward Kirton, Eugene Goltsman, Vasanth Singan, Christopher Hack, Douglas Smith, and Paul Richardson

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

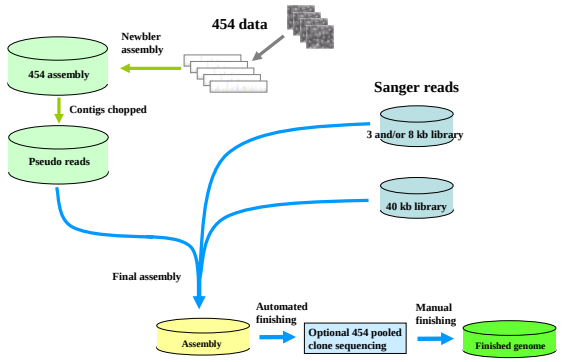
Introduction

The Department of Energy Joint Genome Institute (www.jgi.doe.gov) in Walnut Creek, CA is a high throughput DNA sequencing facility with a current throughput of approximately 3 billion Sanger base pairs per month. A major effort at JGI is the sequencing of microbial genomes of relevance to the DOE missions of carbon sequestration, bioremediation and energy production. The JGI Microbial Program and Community Sequencing Program together are responsible for the generation of sequencing data for over 400 microbial genomes. At the traditional Sanger sequencing side, JGI is running about 70 ABI sequencers on a 24/7 schedule and about 40 GE MegaBACE 4500 sequencers on a 24/5 schedule. JGI currently runs 2 Roche's GS20 instruments to supplement our traditional Sanger sequencing. Our current whole genome shotgun sequencing strategy is to sequence 3kb and/or 8kb shotgun libraries to a combined 4-8x draft coverage and to sequence fosmid ends to 1x sequence coverage with Sanger sequencing and to supplement that with 12-25x coverage with 454 sequencing platform depending on the sizes of the genomes.

For new microbial genomes we initiated, 454 sequencing was carried out at the same time the shotgun cloning for Sanger sequencing started. 454 sequencing data was used to profile the genomes for G/C content, genome sizes and other features of the genomes. For existing microbial genome projects for which the Sanger sequencing data has already been generated, we have been adding 454 sequencing coverage at the finishing stage. 454 sequencing data was assembled by default Newbler assembler software package from 454 Life Sciences. The Newbler contigs were then fragmented and the quality and coverage information of the contigs was captured by in-house developed software tool packages. The fragmentation strategy we currently use is to cut the Newbler contigs into 750 bp fragments with 100 bp overlap. The overlapping fragments from Newbler contigs were finally assembled with Sanger sequencing data using the assembler(s) of our choice. The gaps and low quality areas in the final assembly were manually sequenced to JGI defined quality standard. At this point, the genome is ready for analysis and annotation.

454 sequencing technology is also used more directly in gap closure stage of the microbial genome whole genome shotgun sequencing. Gap spanning clones from multiple genomes were pooled together and the resulting DNA was subjected to 454 sequencing. The Newbler assembly results from pooled clone sequencing were added to final genome assemblies to fill the gaps.

Strategy



Acknowledgements

Finishing Group at JGI
Alla Lapidus
Steve Lowry

Informatics Group at JGI
Stephen Trong
Harris Shpiro

Collaborators
All genomic DNA was supplied by our collaborators
Halothermothrix orenii DSM 9562 — Philips Hugenholtz, Joint Genome Institute
Prochlorococcus marinus MIT9215 — Sallie Chisholm, Massachusetts Institute of Technology
Prochlorococcus marinus NATL2A — Sallie Chisholm, Massachusetts Institute of Technology
Lactobacillus reuteri JCM1112 — Gerald Tannock, University of Otago
Methanocorpusculum labreanum Z — Carl Woese, University of Illinois at Urbana-Champaign
Bifidobacterium longum infantis — David Mills, University of California at Davis
Desulfotomaculum like organism — Terry Hazen, Lawrence Livermore National Laboratory

Feng Chen, Ph.D.
US DOE Joint Genome Institute
2800 Mitchell Dr., B400
Walnut Creek, CA 94598
(925) 296-5733
fchen@lbl.gov
LBNL-61726 Poster

Results

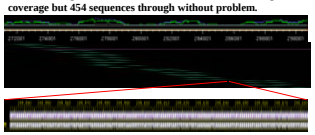
JGI has sequenced over 30 microbial genomes with combined Sanger and 454 sequencing data. Six of these genomes have been chosen for this study. They vary in genome sizes and more importantly, in G/C contents, from 38% to 62%. Sequencing data from one 454 run for each organism was added to different Sanger sequencing coverage. Sanger coverage ranges from 4x for Prochlorococcus MIT9215 to 15x for Lactobacillus reuteri. Lactobacillus reuteri with high A/T content showed strong cloning bias for Sanger sequencing and 16x 454 data didn't show the same effectiveness as with other genomes. For all other genomes, number of gaps tremendously decreased, at least by 80%, with 11x to 24x 454 sequencing data. The gaps from both platforms are not distributed evenly throughout the genomes and the size of gaps varies.

Organism	GC %	Genome size	# of Gaps w/o 454	# of Gaps with 454	454 Depth	454-only bp	Sanger-only bp
Halothermothrix orenii DSM 9562	38	2.58	37	7	15x	21801	60837
Prochlorococcus marinus MIT9215	39	1.74	141	2	15x	26333	42722
Lactobacillus reuteri JCM 1112	39	2.03	42	15	16x	44701	145883
Methanocorpusculum labreanum Z	50	1.81	11	2	22x	426	17769
Bifidobacterium longum infatis	60	2.83	9	2	11x	348	71061
Desulfotomaculum like organism	62	2.35	20	1	24x	14122	111327

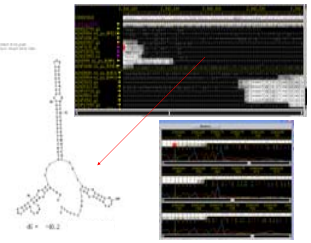
Quality of 454 run for each genome was individually assessed by comparing 454 consensus sequences to aligned high quality Sanger consensus sequences. Error prone ends of the Newbler contigs were trimmed and excluded from the comparison. Errors from Newbler mis-assembling were also excluded in the calculation.

ORGANISM	Miscall	Insertion	Deletion	HMP deletion	HMP insertion	HMPinDel	Incomplete Extension	Other	TOTAL	Aligned	% Aligned	ACCURACY	Q-SCORE
Halothermothrix_orenii_DSM9562.	22	122	88	1123	764	0	204	0	2323	2536678	98.46	99.90842	30.4
Prochlorococcus_marinus_MIT9215	1	19	525	238	0	431	2	9	1225	1736228	99.81	99.92944	31.5
Lactobacillus reuteri JCM-1112	384	36	80	202	73	0	80	7	862	1888286	92.91	99.95435	33.4
Methanocorpusculum_labreanum_Z	40	3	11	30	13	0	4	2	103	1788704	99.08	99.99424	42.4
Bifidobacterium_longum_infantis	59	16	16	21	10	0	6	4	132	2775413	97.98	99.99524	43.2
Desulfotomaculum-like_organism	21	22	24	503	137	0	25	0	732	2272877	96.74	99.96779	34.9
Prochlorococcus_marinus_NATL2A	13	87	38	448	224	0	211	0	1021	1825430	98.55	99.94407	32.5

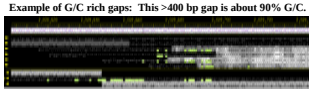
Example of A/T rich gaps:
This >1 kb region is more than 80% A/T, there is no Sanger coverage but 454 sequences through without problem.



Example of secondary structure hard-stop gaps:

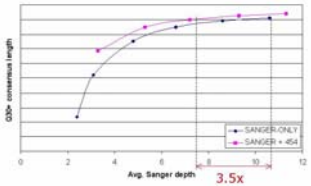


Example of G/C rich gaps: This >400 bp gap is about 90% G/C.

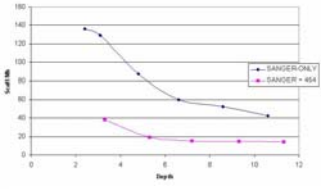


Comparison of Sanger only assembly and 454+Sanger assembly

Consensus length of L. reuteri



Uncaptured gaps of L. reuteri



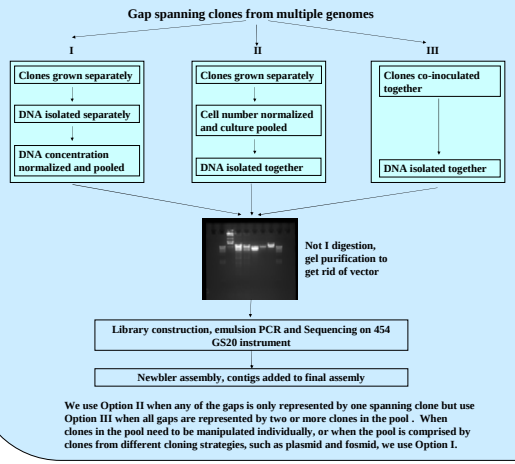
Discussion and Conclusions

Combination of 454 sequencing and traditional Sanger sequencing technology is an efficient strategy for microbial whole genome shotgun sequencing. The different technologies have different biases in their processes and they seem to be complementary in WGS sequencing to provide more random and even coverage than either one of the technologies can do alone. This study with limited number of genomes showed that 454 sequencing can effectively cover more than 80% of the regions which were not covered by Sanger sequencing even when the depth of Sanger sequencing has already been high (up to 13x). Traditional gap closure methods involve extensive manual lab work and long turn-around time. 454 sequencing can significantly decrease both of them. With 454 sequencing technology, we are able to decrease the coverage of Sanger sequencing by about 4x and still achieve similar assembly quality.

The quality of the 454 sequencing is high from the genomes included in this study when we excluded the ends of the Newbler contigs and the mis-assembled regions of Newbler assembly. Currently, our finishing group is still validating all 454 read only regions with Sanger sequencing. A new strategy is under the discussion to eliminate some of this polishing process when we have the consensus of the confidence level of the quality of the 454 read only regions.

Sequencing Pooled Gap-Spanning Clones Using 454 Platform

Methods



We use Option II when any of the gaps is only represented by one spanning clone but use Option III when all gaps are represented by two or more clones in the pool . When clones in the pool need to be manipulated individually, or when the pool is comprised by clones from different cloning strategies, such as plasmid and fosmid, we use Option I.

Results

Control fosmid:
One of the fosmids in the pool was actually fully sequenced by Sanger shotgun sequencing, and included as a reference. The following depiction of the BLAST alignment results shows that this fosmid was reasonably well covered by 454 contigs. The uncovered region of ~3 kb may either represent an area not covered by 454 sequencing or a misassembly in the Sanger reference sequence.

Example genome:
All gaps were represented by 2 or more spanning clones for this experiment. Thirty-two fosmid clones were pooled and the expected average 454 sequencing coverage for each clone is at about 30x. All 13 gaps were successfully closed by 454 sequencing.

Draft assembly | After auto finishing | After 454 pooled clone sequencing

Example of one actual fosmid:

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396.